Am. J. Hum. Genet. 60:957–964, 1997

Ethnic-Affiliation Estimation by Use of Population-Specific DNA Markers

Mark D. Shriver,[1] Michael W. Smith,[2] Li Jin,[3] Amy Marcini,[1] Joshua M. Akey,[1] Ranjan Deka,[1] and Robert E. Ferrell[1]

[1]Department of Human Genetics, University of Pittsburgh, Pittsburgh; [2]Intramural Research Support Program, SAIC–Frederick, National Cancer Institute, Frederick Cancer Research and Development Center, Frederick, MD; and [3]Department of Genetics, Stanford School of Medicine, Stanford

Summary

During the past 10 years, DNA analysis has revolutionized the determination of identity in a forensic context. Statements about the biological identity of two human DNA samples now can be made with complete confidence. Although DNA markers are very powerful for distinguishing among individuals, most offer little power to distinguish ethnicity or to support any statement about the physical characteristics of an individual. Through a search of the literature and of unpublished data on allele frequencies we have identified a panel of population-specific genetic markers that enable robust ethnic-affiliation estimation for major U.S. resident populations. In this report, we identify these loci and present their levels of allele-frequency differential between ethnically defined samples, and we demonstrate, using log-likelihood analysis, that this panel of markers provides significant statistical power for ethnic-affiliation estimation. In addition to their use in forensic ethnic-affiliation estimation, population-specific genetic markers are very useful in both population- and individual-level admixture estimation and in mapping genes by use of the linkage disequilibrium created when populations hybridize.

Introduction

It is well documented that the majority (80%–90%) of the genetic variation among humans is interindividual and that only 10%–20% of the total variation is due to population differences (e.g., see Nei 1987; Cavalli-Sforza et al. 1994; Deka et al. 1995). It also has been widely observed that most populations share alleles at any given locus and that those alleles that are most frequent in one population are also found at high frequency in other populations. There are few classical (blood group, protein, and immunological) or DNA genetic markers that have been demonstrated either to be population specific or to have large frequency differentials among geographically and ethnically defined populations (Roychoudhury and Nei 1988; Cavalli-Sforza et al. 1994; Dean et al. 1994). The high degree of genetic similarity reflects the recent divergence of the human species into continental groups. A common recent origin for all human populations is supported by both archeological and genetic evidence. Despite this apparent lack of unique genetic markers, there are marked physical and physiological differences among human racial groups, which, it is presumed, reflect long-term adaptation to unique ecological conditions, random genetic drift, and sexual selection.

The idea that there are genetic markers that exist in one population and not in others was first presented by Neel (1973), who referred to these markers as "private" and used them to estimate mutation rates. Reed (1973) used the term "ideal" (in reference to their utility in individual-admixture estimation) to describe hypothetical genetic marker loci at which different alleles are fixed in different populations. Chakraborty et al. (1991) called those variants that are found in only one population "unique alleles." The most useful unique alleles for forensics, admixture, or mapping studies are those which have the largest allele-frequency differences among populations (Reed 1973; Chakraborty et al. 1992; Stephens et al. 1994). We use the designation "population-specific alleles"(PSAs) to describe genetic markers with large allele-frequency differentials. For a biallelic marker the frequency differential (δ) is equal to $p_X - p_Y$, which is equal to $q_Y - q_X$, where p_X and p_Y are the frequencies of one allele in populations X and Y and q_X and q_Y are the frequencies of the other. Although the PSA marker alleles referred to may not be absolutely restricted to any population, the name underscores the primary uniqueness and utility of these markers. Median δ levels for biallelic loci among major ethnic groups are 15%–20%, and the vast majority (>95%) of arbitrarily identified biallelic genetic markers have δ

Received October 24, 1996; accepted for publication January 23, 1997.

Address for correspondence and reprints: Dr. Mark D. Shriver, Department of Human Genetics, University of Pittsburgh, 130 De Soto Street, Pittsburgh, PA 15235.

*Present affiliation: Human Genetics Center, University of Texas–Houston.

© 1997 by The American Society of Human Genetics. All rights reserved.
0002-9297/97/6004-0026$02.00

<50% (Dean et al. 1994). We thus propose that a "large" frequency differential is >50% and that PSAs are those markers that demonstrate such frequency differences between any two major geographically or ethnically defined populations. A definition of PSAs is also possible when a locus has more than two alleles. When there are more than two alleles, δ is calculated as the summation of all those individual δ values that have like sign when allele frequencies for two populations are subtracted.

A selection of PSAs would be useful in several areas of human genetics, physical anthropology, and forensic science. In forensic anthropology, the determination of the race or ethnic affiliation of an individual is difficult, without access to skin and hair samples. The best racial estimates are achieved by use of a number of measurements of lengths and angles from the skull and several large bones, which are used to estimate coefficients of discriminant functions. Using this approach, researchers are able to correctly categorize 80%–90% of the individuals in the sample used to estimate the coefficients (Dibennardo and Taylor 1983; Iscan 1983). The accuracy of classification has been observed to be lower when samples other than the ones used to estimate the discriminant coefficients are analyzed. In addition, since the large U.S. skeleton collections were established several decades ago, the appropriateness of estimates based on these collections has been questioned (Dibennardo and Taylor 1983). Since ethnic-affiliation estimation (EAE), in turn, is used to estimate other characteristics, such as gender, age, and height, precise EAEs will enhance, in more than one way, the accuracy of the physical description. For example, age at time of death is estimated by use of measurements of the long bones and by standard formulas. One important factor in these formulas is ethnicity, since the relationship between age and height is not the same across ethnically defined groups.

The precision of estimates of admixture is directly dependent on the δ level of the markers used (Chakraborty et al. 1991). PSAs are thus the ideal markers for accurate population-admixture estimation and will make individual-admixture estimation feasible. Hybrid populations, such as the African Americans and Hispanic Americans, also can be a valuable resource for mapping disease genes. This method is called "*m*apping by *a*dmixture *l*inkage *d*isequilibrium" (MALD) and uses the linkage disequilibrium created when divergent populations hybridize to localize genes that predispose to disease (Chakraborty and Weiss 1988; Briscoe et al. 1994; Stephens et al. 1994). Since the magnitude of the admixture linkage disequilibrium is proportional to the δ level of the marker loci used for MALD mapping, PSAs are the preferred markers for this method.

Methods

Single-locus average log-likelihood determinations were made by first calculating the estimated genotype frequencies by use of observed allele frequencies for all possible genotypes in the first population. The frequency of each genotype was then calculated for the second population by use of the allele frequencies for that population. If a particular allele was not observed in the second population, the frequency was set to $1/(2n + 1)$, where n is the sample size that was typed to determine allele frequencies. In other words, an allele not found in a sample is assumed to be the next allele to be observed, thus preventing undefined log-likelihood ratios due to division by zero. The individual genotype log-likelihood ratios can be expressed as

$$LLR_{Axy} = \log_{10}\left(\frac{2a_x a_y}{2b_x b_y}\right)$$

and

$$LLR_{Bxy} = \log_{10}\left(\frac{2b_x b_y}{2a_x a_y}\right),$$

when $x \neq y$ or

$$LLR_{Axy} = \log_{10}\left(\frac{a_x a_y}{b_x b_y}\right)$$

and

$$LLR_{Bxy} = \log_{10}\left(\frac{b_x b_y}{a_x a_y}\right),$$

when $x = y$ and a_x and a_y are the allele frequencies of alleles x and y in population A and b_x and b_y are the frequencies of alleles x and y in population B. The average single-locus EAE log-likelihood ratio is then

$$\sum_{x=1}^{k} \sum_{y=x}^{k} (\tfrac{1}{2} P_{Axy} LLR_{Axy} + \tfrac{1}{2} P_{Bxy} LLR_{Bxy}),$$

where k is the number of alleles at the locus and P_{Axy} and P_{Bxy} are the genotype frequencies for genotype xy in populations A and B, respectively.

The expected value of the EAE log-likelihood ratio across multiple loci can be calculated simply as the sum of the average single-locus log-likelihood ratios. Given the exponential relationship between the number of loci and the number of multilocus genotypes, it is not possible to determine directly the distribution of log-likeli-

hood levels when more than a few loci are used. Instead, we have used computer simulations to calculate the average log-likelihood ratio and to create the distribution of log-likelihood levels when several loci were considered simultaneously. By use of random numbers, an individual was created on the basis of the allele frequencies in the first population and under the assumption of linkage equilibrium among all alleles. Then the frequency of this multilocus genotype was calculated in the other population, again with the stipulation that alleles not observed have frequencies of $1/(2n + 1)$, where n is the sample size. The log is then taken on the ratio of the two multilocus genotype-frequency estimates. Each multilocus log-likelihood distribution presented is the result of 100,000 simulated individuals.

DNA was isolated from frozen buffy coats or cell pellets and was used as a template in PCR-based (Saiki et al. 1988) restriction-site or microsatellite PCR genotyping reactions. For restriction-site polymorphism typing, standard PCR using primers 20–25 bp in length was done in a 25-μl volume. Typically, restriction enzymes (1–5 U) were added directly to an aliquot of the PCR reaction, without modification of the reaction buffer, and were allowed to digest overnight at the prescribed temperature. Restriction fragments then were separated by either agarose-gel electrophoresis or PAGE, depending on the predicted fragment sizes. Microsatellite markers were amplified by use of either isotopically labeled or fluorescently labeled oligonucleotide primers and were run on denaturing polyacrylamide gels to separate alleles by size. Readers interested in primer sequences and reaction conditions are referred to the Genome Database (GDB; www.gdb.org); entries are referenced in the text below or in tables 1 and 2. Details on the populations analyzed have been presented elsewhere (see Bowcock et al. 1994; Dean et al. 1994; Deka et al. 1995).

Results

Dimorphic PSAs

We first searched both the literature and our own unpublished data, for dimorphic and microsatellite PSA markers that would be useful in EAE of U.S. resident populations—namely, Africans, Europeans, Amerindians, European Americans, African Americans, and Hispanic Americans. There have been five large surveys of RFLP loci detected by use of Southern blot techniques (Bowcock et al. 1987; Roychoudhury and Nei 1988; Bowcock et al. 1991; Kidd et al. 1991; Dean et al. 1994) and one population survey of RFLP loci by use of PCR-based restriction-site polymorphisms (Jorde et al. 1995), in both of which, for some of these populations, frequencies were reported, for a total of 565 loci comparisons. Only 28 (5%) of these 565 loci/pair combinations of RFLP loci showed levels of $\delta > .5$ (the cutoff level for PSAs). However, most of these markers have not been developed into PCR-based genotyping assays, and thus they are of limited use in most modern molecular-genetics laboratories.

From the literature, we identified a group of six PCR-based dimorphic genetic markers that have been typed in ethnically defined populations and that have been reported to have large allele-frequency differentials. We were able to verify that all six of these markers—namely, FY-null (GDB designation 728415), RB2300 (GDB designation 155206), LPL (GDB designation 157022), CKMM (GDB designation 156902), PV92 (Batzer et al. 1994), and DRD2 (K. K. Kidd, personal communication)—have δ levels high enough to make them useful for EAE. Three of these loci (FY-null, RB2300, and LPL) have high African/European and African/Amerindian δ levels, and three (CKMM, PV92, and DRD2) have high Amerindian/African and Amerindian/European frequency differentials. The allele frequencies for these six loci in five world and U.S. populations are presented graphically in figure 1. It is clear that there are large allele-frequency differences among populations and that the two hybrid populations have allele frequencies intermediate between those of the parental populations. For example, the frequency of the FY-null allele is 100% in our African sample, very low in Europeans (and all other populations), and 80% in African Americans.

Hypervariable Microsatellite PSAs

Hypervariable microsatellites, short tandem arrays of 2–6-bp repeat units, are the preferred DNA markers in many areas of genetic research, ranging from studies of microevolution to gene mapping and identity analysis. Microsatellites are useful because they demonstrate very high levels of heterozygosity (usually >.70), which are due to elevated mutation rates (~1 mutation/1,000 gametes; Weber and Wong 1993). There is good evidence that replication slippage is the mechanism of mutation and that changes in array length proceed in a stepwise manner, with mutant alleles becoming either one repeat unit larger or one repeat unit smaller (Levinson and Gutman 1987; Shriver et al. 1993; Valdes et al. 1993; DiRienzo et al. 1994). Those characteristics—high heterozygosity and large numbers of alleles—which make microsatellites ideal markers for most genetic research, do not directly affect their utility in EAE, admixture estimation, or MALD. The primary determinant of their utility in these applications is, as with diallelic markers, the frequency differential between populations. Indirectly, the higher mutation rate at microsatellite loci leads to more alleles—and thus to rapid divergence of these loci—because each allele is subject to random changes in frequency. Although several thousand micro-

Table 1

δ And Average Log-Likelihood Levels for 20 African PSA Loci

	δ; Average Log-Likelihood Level		
Locus[a]	African American/European American	African American/Hispanic American	Hispanic American/European American
FY-null[b]	.783; 1.858	.781; 1.818	.002; .001
D7S657	.745; 1.276	.556; .752	.223; .138
D5S421	.714; 1.094	.407; .391	.357; .280
D14S72	.654; 1.085	.718; 1.561	.175; .128
D6S276	.652; 1.029	.595; .962	.218; .170
D9S175	.636; .992	.609; .914	.267; .216
D2S338	.633; .903	.528; .658	.400; .443
D2S117	.632; .910	.286; .279	.432; .441
RB2300[b]	.617; .837	.683; 1.033	.067; .010
D5S410	.595; .899	.548; .761	.220; .109
DRPLA	.594; .978	.452; .452	.214; .190
D4S413	.592; .722	.630; .948	.324; .337
D7S640	.582; .794	.536; .635	.452; .407
D8S272	.575; .737	.312; .257	.465; .552
D1S218	.563; .746	.450; .500	.225; .186
D12S101	.556; .634	.444; .437	.204; .178
D8S284	.500; .636	.390; .432	.198; .153
D8S284	.500; .636	.390; .432	.198; .153
LPL[b]	.483; .532	.500; .565	.017; .001
D15S120	.449; .498	.381; .354	.297; .229

[a] Microsatellites are listed by use of the DS nomenclature. Primer sequences and PCR conditions can be found in the Genome Database (GDB; www.gdb.org). Conditions for FY-null, RB2300, and LPL are referenced in the text.
[b] Dimorphic loci.

Table 2

δ And Average Log Likelihood Levels for 20 Hispanic PSA Loci

	δ; Average Log-Likelihood Level		
	Hispanic American/European American	Hispanic American/African American	African American/European American
D1S255	.524; .632	.262; .169	.500; .478
D7S530	.489; .635	.750; .667	.364; .359
D8S272	.465; .552	.312; .257	.575; .737
D7S640	.452; .407	.536; .635	.582; .794
D3S1211	.441; .599	.460; .565	.302; .315
D15S131	.438; .449	.357; .266	.319; .342
D2S396	.436; .414	.336; .268	.230; .170
D2S338	.432; .441	.286; .279	.633; .908
D5S406	.432; .525	.486; .646	.516; .599
D3S1297	.417; .547	.425; .455	.460; .643
D1S249	.415; .389	.333; .253	.281; .194
D7S550	.415; .411	.571; .832	.605; .879
D9S156	.405; .321	.203; .139	.367; .306
D9S175	.400; .443	.528; .658	.636; .992
D3S1278	.397; .337	.563; .859	.392; .402
D1S502	.395; .414	.710; .803	.711; .729
D8S279	.387; .505	.315; .280	.382; .332
D12S83	.383; .373	.468; .525	.488; .549
D3S1289	.380; .331	.202; .223	.401; .498
D11S937	.380; .326	.532; .603	.489; .541

[a] Microsatellites are listed by use of the DS nomenclature. Primer sequences and PCR conditions can be found in the Genome Database (GDB; www.gdb.org).

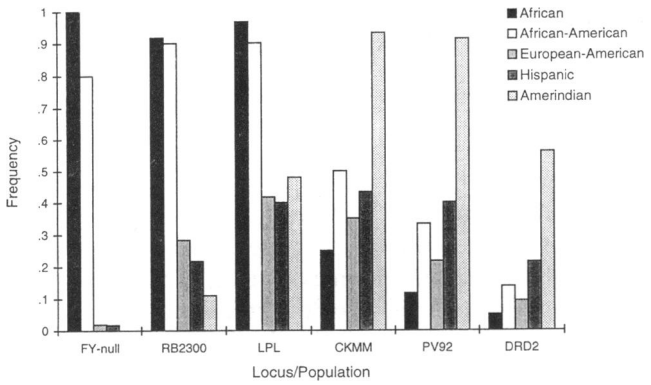

Figure 1 Frequencies of six dimorphic-PSAs in five populations. The sample size for each of four populations (Africans, African Americans, European Americans, and Hispanics) is 30, whereas, for the Amerindian group (Pehuenche), we typed 50 individuals.

satellite loci have been isolated and mapped in the human genome, relatively few data on their allele-frequency distributions across populations are available. In the course of this work we have inspected the levels of frequency differential for >1,000 loci, including both unpublished data generated in our laboratories and data from three published studies (Bowcock et al. 1994; Deka et al. 1995; and Jorde et al. 1995).

Shown in figure 2*A* and *B* are examples of microsatellite PSA markers identified in these surveys. In figure 2*A* we show the allele-frequency distributions of an African/European PSA, D7S657, that has a δ level of .745. Large differences in the frequencies of particular alleles are evident (e.g., alleles 248 and 262). There are also some less frequent alleles, which were found in only one of the samples. Figure 2*B* shows the allele-frequency distributions of the Hispanic/European PSA, D1S255. This locus demonstrates a δ level of .524 between European Americans and Hispanic Americans, most of which is due to three alleles (alleles 73, 75, and 79).

In tables 1 and 2 we present lists of the PSA loci identified in these surveys which are best suited to enable confident EAE in U.S. populations. These loci are the subset of those PSAs found that will provide the most statistical power for EAE in U.S. populations; they do not comprise a complete list (such a presentation is beyond the scope of this report and will be presented elsewhere [M. W. Smith, R. Deka, and L. Jin, unpublished data]). We have based the selection of loci for this list on the levels of allele-frequency differential among the ethnically classified individuals typed in our laboratories. All of these loci are genotyped by use of PCR, and the majority are microsatellite polymorphisms. Part of our selection criterion was that the identified loci should be widely dispersed across the genome, to minimize the effects that linkage disequilibrium would have on the computation of the log-likelihood ratios. Since there may be substantial uncertainty in allele-frequency estimates made by use of small sample sizes, and because, for small sample sizes, the estimation of δ is biased in an upward direction (data not shown), we have also excluded those loci that have been typed in samples of <20 individuals. Also presented in tables 1 and 2 are the pairwise δ levels and average single-locus log-likelihood levels for these loci, for the major U.S. population groups. These calculations are all based on observed allele frequencies for the primary U.S. resident populations (namely, European American, African American, and Hispanic American) and not on the parental or unadmixed population frequencies.

To characterize the informativeness of the PSA panels presented in tables 1 and 2, we have determined the distribution of multilocus log-likelihood levels. These

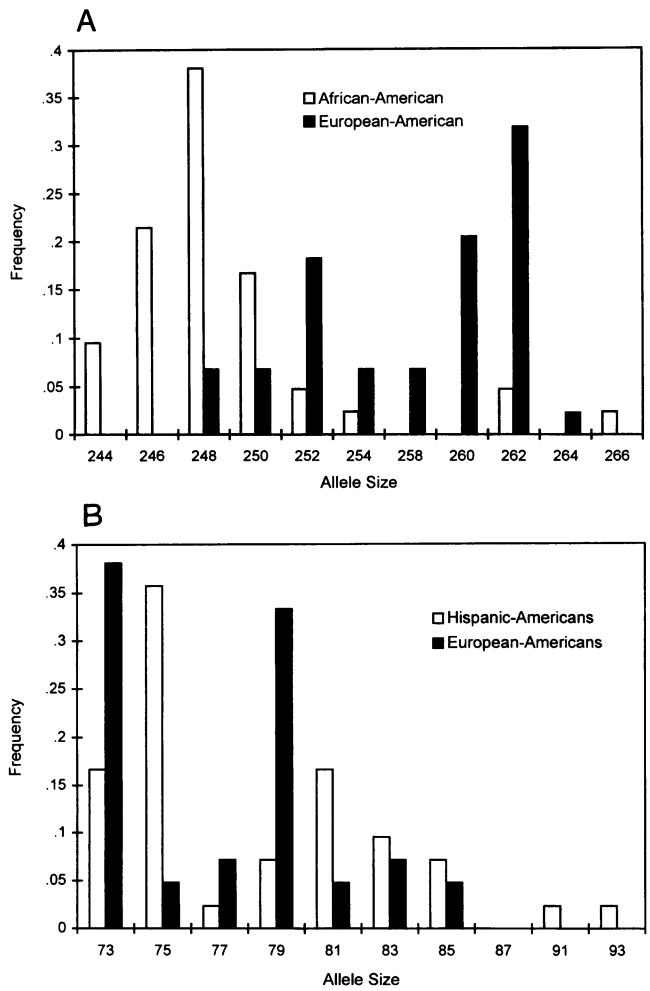

Figure 2 Allele-frequency distributions for two microsatellite-PSA loci. *A*, Frequencies of the best African microsatellite-PSA D7S657 in an African American sample (n = 21) and in an European American sample (n = 22). *B*, Frequencies of the best Hispanic microsatellite D1S255 in a U.S. Hispanic sample (n = 21) and in a U.S. European American sample (n = 21).

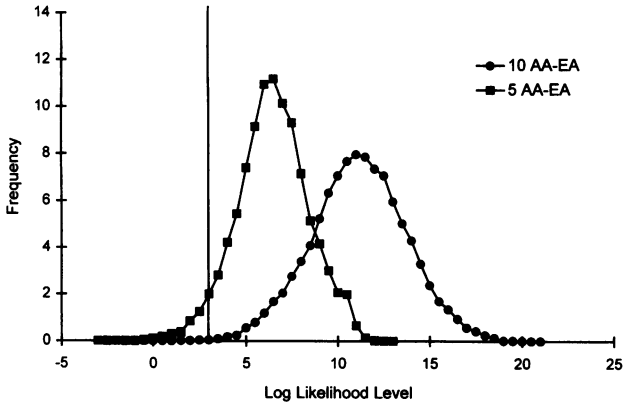

Figure 3 Distribution of multilocus average log-likelihood levels for African American/European American EAE. Two lines are shown, representing the results of EAE for the 5 best (*blackened squares*) and 10 best (*blackened circles*) African American PSAs from table 1. Each line is the probability distribution of the multilocus log-likelihood level, based on 10,000 simulated individuals. A vertical line intersects the *X*-axis at a log-likelihood level of 3.0.

results are shown in figures 3 and 4. Figure 3 shows the distribution of multilocus log likelihood for African Americans versus European Americans, and figure 4 shows the same for Hispanic Americans versus European Americans. As expected, given the substantial European genetic component in Hispanic populations, Hispanic American versus African American log-likelihood distributions were very similar to the African American versus European American distributions and therefore are not shown. Two curves are shown in figure 3, the results of using the 5 best PSAs (the first five markers from table 1) and the 10 best PSAs (the first 10 markers from table 1). Note that a vertical line intersects the *X*-axis at a log-likelihood level of 3 (odds of 1:1,000, a commonly accepted level of scientific confidence). For African Americans, the average multilocus log-likelihood level is 6.5 when the 5 best PSAs are used and is 11.1 when the 10 best PSAs are used. When the panel of the 5 best PSAs are used, 96.0% of individuals have log-likelihood levels >3.0; when the entire panel of 10 is used, <0.01% show log likelihoods <3.0.

The Hispanic American distributions of log likelihood are presented in figure 4. Three curves are shown: one for the 5 best, one for the best 10, and one for the best 20 PSAs from table 2. As expected, given the higher European genetic contribution to the Hispanic American population than to the African American population, the log-likelihood levels for these comparisons are lower. The average log likelihood for the panel of 5 PSAs is 2.9, and 46.8% of individuals yield a log-likelihood level >3.0; for the best 10 and 20, the average log-likelihood levels are 5.2 and 9.1, respectively, and the proportions of observations for which the log-likelihood level was >3.0 were 87.4% and 99.0%, respectively.

Discussion

The principal goal of this project was to identify a set of genetic markers that would allow the confident determination of ethnicity, for use in a forensic setting. Although ethnic affiliation is often clearly evident on gross observation, many of the traits that allow these distinctions are superficial. Ethnic classification is much more difficult when one has only skeletal remains or a sample of blood to examine. This is primarily due to the facts that human populations share a very recent common ancestry and that the majority of the total genetic variation is due to differences within populations and not to differences between them. Thus, alleles found in one population often are found in others, and those which are most frequent in one are also frequent in others. Additionally, the large-scale hybridization or admixture of populations that has occurred in the United States has acted to obscure the genetic differences among resident populations. Despite these factors, we have shown that it is possible to identify a collection of genetic markers that are distinctive enough to allow confident genetic EAE. The principal characteristic of these markers, which we call "PSAs," is that they have substantial allele-frequency differences (high δ levels) between populations.

Through a search of the literature and of unpublished data from our laboratories, we have compiled two selections of PSA markers, one showing high δ levels between African Americans and European Americans and the

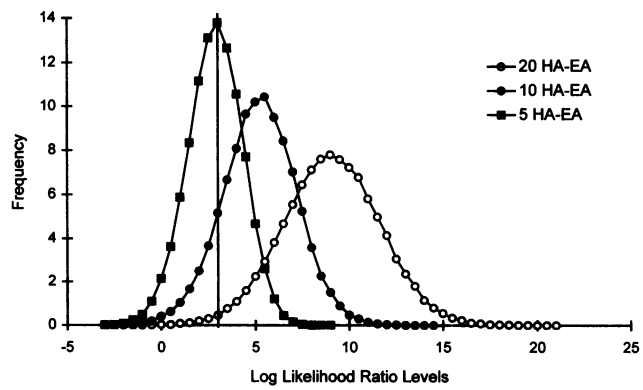

Figure 4 Distribution of multilocus average log-likelihood levels for Hispanic American/European American EAE. Three lines are shown, representing the results of EAE for the 5 best (*blackened squares*), 10 best (*blackened circles*), and 20 best (*unblackened circles*) Hispanic American PSAs from table 2. Each line is the probability distribution of the multilocus log-likelihood level, based on 10,000 simulated individuals. A vertical line intersects the *X*-axis at a log-likelihood level of 3.0.

other showing high δ levels between European Americans and Hispanic Americans. For inclusion within this list, loci were selected on the basis of high δ and high log-likelihood levels. In addition, we have included only loci for which PCR-based assays are available. We have presented a list of 20 African PSAs and 20 Hispanic PSAs arranged in descending order of δ level—and, thus, in descending level of informativeness for EAE. It should be noted that the markers on this list need to be typed in larger samples from different parts of the country, both to have more accurate allele-frequency estimates and to identify the most efficient set for EAE. We show, using simulation-based multilocus log-likelihood determinations, that a subset of these markers provide for confident EAE among these three populations. Multilocus EAE log likelihoods between African Americans and either European Americans or Hispanic Americans are much higher than those between Hispanic Americans and either European Americans or African Americans. This observation is consistent with data from other genetic studies, which show that the earliest branch in the human phylogenetic tree is between Africans and non-Africans and that the European genetic component is ~25% in contemporary African American populations (Chakraborty et al. 1991) and ~60% in contemporary Hispanic populations (Hanis et al. 1991).

Connor and Stoneking (1994) have applied mitochondrial genetic variation to the estimation of ethnicity. This group used a logistic-regression model to predict ethnicity on the basis of sequence-specific oligonucleotide data and demonstrated that, with 23 probes, they are able, 65% of the time, to correctly predict ethnic group. This is a significant level of discrimination, especially when one considers that the mitochondrial genome is a single genetic locus, albeit a highly informative locus. Both the limited time since divergence and ancient and historical admixture prohibit the practical application of a single locus to EAE. Many informative loci, possibly including Y-chromosome and mtDNA markers, are needed for confident EAE, individual-admixture estimation, or group-admixture estimation. Given both the recent divergence of human populations and the strong tendency for populations to hybridize, the variance of estimates based on one of a few loci will often be too high for confident statements to be made.

We have demonstrated that it is possible to identify a panel of dimorphic and microsatellite genetic markers that will allow confident EAE in African Americans, European Americans, and Hispanic Americans. We have focused on these three populations because together they constitute 95% of U.S. residents. Similar sets of markers could be developed for the identification of other populations common in the United States, such as Chinese Americans, Native Americans, and Polynesian Americans. In addition, it may prove feasible to estimate individual admixture concurrent with EAE, so that interethnic individuals, first- or second-generation hybrids of one or more populations, could be identified and classified appropriately.

Acknowledgments

This work was supported in part by National Institute of Justice grant 95-IJ-CX-0008 and a grant from the Keck Foundation for Advanced Training in Computational Biology, both to M.D.S.; by National Institutes of Health grant GM-45861, to R.D.; and by National Institutes of Health training grant T32-GS08404, to L.J. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does the mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

References

Batzer MA, Stoneking M, Alegria-Hartman M, Bazan H, Kass DH, Shikh TH, Novick GE, et al (1994) African origin of human-specific polymorphic *Alu* insertions. Proc Natl Acad Sci USA 91:12288–12292

Bowcock AM, Bucci C, Herbert JM, Kidd JR, Kidd KK, Friedlander JS, Cavalli-Sforza LL (1987) Study of 47 DNA markers in five populations from four continents. Gene Geogr 1: 47–64

Bowcock AM, Herbert JM, Mountain JL, Kidd JR, Rodgers J, Kidd KK, Cavalli-Sforza LL (1991) Study of an additional 58 DNA markers in five human populations from four continents. Gene Geogr 5:151–173

Bowcock AM, Ruiz-Linares A, Tomfohrde J, Minch E, Kidd JR, Cavalli-Sforza LL (1994) High resolution of human evolutionary trees with polymorphic microsatellites. Nature 368:455–457

Briscoe D, Stephens JC, O'Brien SJ (1994) Linkage disequilibrium in admixed populations: applications in gene mapping. J Hered 85:59–63

Cavalli-Sforza LL, Menozzi P, Piazza A (1994) The history and geography of human genes. Princeton University Press, Princeton, NJ

Chakraborty R, Kamboh MI, Ferrell RE (1991) "Unique" alleles in admixed populations: a strategy for determining hereditary population differences of disease frequencies. Ethn Dis 1:245–256

Chakraborty R, Kamboh MI, Nwankwo M, Ferrell RE (1992) Caucasian genes in American blacks: new data. Am J Hum Genet 50:145–155

Chakraborty R, Weiss KM (1988) Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. Proc Natl Acad Sci USA 85:9119–9123

Connor A, Stoneking M (1994) Assessing ethnicity from human mitochondrial DNA types determined by hybridization with sequence-specific oligonucleotides. J Forensic Sci 39: 1360–1371

Dean M, Stephens JC, Winkler C, Lomb DA, Ramsburg M,

Boaze R, Stewart C, et al (1994) Polymorphic admixture typing in human ethnic populations. Am J Hum Genet 55: 788–808

Deka R, Jin L, Shriver MD, Yu LM, DeCroo S, Hundrieser J, Bunker CH, et al (1995) Population genetics of dinucleotide $(dC\text{-}dA)_n/(dG\text{-}dT)_n$ polymorphisms in world populations. Am J Hum Genet 56:461–474

Dibennardo R, Taylor JV (1983) Multiple discriminant function analysis of sex and race in the postcranial skeleton. Am J Phys Anthropol 61:305–314

DiRienzo A, Peterson AC, Garza JC, Valdes AM, Slatkin M, Freimer NB (1994) Mutational processes of simple-sequence repeat loci in human populations. Proc Natl Acad Sci USA 91:3166–3170

Hanis CL, Hewett-Emmett D, Bertin TK, Schull WJ (1991) Origins of US Hispanics: implications for diabetes. Diabetes Care 14:618–627

Iscan MY (1983) Assessment of race from the pelvis. Am J Phys Anthropol 62:205–208

Jorde LB, Bamshad MJ, Watkins WS, Zenger R, Fraley AE, Krakowiak PA, Carpenter KD, et al (1995) Origins and affinities of modern humans: a comparison of mitochondrial and nuclear genetic data. Am J Hum Genet 57:523–538

Kidd JR, Black FL, Weiss KM, Balaza I, Kidd KK (1991) Studies of three Amerindian populations using nuclear DNA polymorphisms. Hum Biol 63:775–794

Levinson G, Gutman GA (1987) Slipped-strand mispairing: a major mechanism for DNA sequence evolution. Mol Biol Evol 4:203–221

Neel JV (1973) "Private" genetic variants and the frequency of mutation among South American Indians. Proc Natl Acad Sci USA 70:3311–3315

Nei, M (1987) Molecular population genetics. Columbia University Press, New York

Reed TE (1973) Number of gene loci required for accurate estimation of ancestral population proportions in individual human hybrids. Nature 244:575–576

Roychoudhury AK, Nei M (1988) Human polymorphic genes: world distribution. Oxford University Press, New York

Saiki RK, Gelfand DH, Stoffel S, Scharf SJ, Higuchi R, Horn GT, Mullis KB, et al (1988) Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. Science 239:487–491

Shriver MD, Jin L, Chakraborty R, Boerwinkle E (1993) VNTR allele frequency distributions under a stepwise mutation model: a computer simulation approach. Genetics 134: 983–993

Stephens JC, Briscoe D, O'Brien S (1994) Mapping by admixture linkage disequilibrium in human populations: limits and guidelines. Am J Hum Genet 55:809–824

Valdes AM, Slatkin M, Freimer NB (1993) Allele frequencies at microsatellite loci: the stepwise mutation model revisited. Genetics 133:737–749

Weber JL, Wong C (1993) Mutation of human short tandem repeats. Hum Mol Genet 8:1123–1128